# JSON Vignette

Accidental Drug Deaths 2012-2018

*Shaleni Kovach*

*10/9/2019*

## Data Type - JSON

This vignette covers JSON data. JSON stands for JavaScript Object Notation, and is commonly used to transfer data as it easy for both humans and machines to parse. JSON is language agnostic, which also makes it an ideal medium for data transfer. JSON is essentially a collection of name-value pairs known as objects. The standard notation for an object is: { name: value }, with the {} marking the beginning and end of an object.

Objects are not limited to one set of names/values, and can also be nested. For example, an object can take on the form:

{
> name1: value1,
> name2: value2
}
or with nesting:
{
> name1: value1,
> name2:
> > { name3: value2 }
}

These objects can further be stored and sent in arrays. If we had data on some students and wanted to send/recieve that data as JSON, one possible way that could be formatted would be: [ student1: { hw1: { part1: score, part2: score }, exam1: { question1: score, question2: score } }, student2: { hw1: { part1: score, part2: score }, exam1: { question1: score, question2: score } },]

and so on and so forth.

## Reading in JSON files

There are three main packages for reading in JSON data in R. These are:

- rjson
- jsonlite
- rjsonio

The JSON file is accessible via an API, and `jsonlite` will be used to read in the data. The `jsonlite` package was chosen due to ease of finding documentation on it, and the format in which it brings in the data - when using the `fromJSON` function, data is stored automatically as a tibble.

## Dataset

This example will use data from a JSON on Crimes by County and Agency from 1990. This dataset is available through the New York OIT Website (website link here).

This dataset contains counts of various categories of crimes by year reported by different counties and agencies in New York. There are two general categories of crime, violent and property, which fall into the subcategories of murder, rape, robbery, aggravated assault and burglary, larceny, and motor vehicle theft respectively.

The data contains 20,000 rows and 13 columns, where each row is report of counts by county and reporting agency.

# EDA

Reading in the data from the NY Data API

```
data <- fromJSON("https://data.ny.gov/resource/ca8h-8gjq.json")
data$county <- as.factor(data$county)
data$agency <- as.factor(data$agency)

head(data)
```

```
##   county                     agency year months_reported
## 1 Albany             Albany City PD 2018              12
## 2 Albany      Albany County Park PD 2018               7
## 3 Albany      Albany County Sheriff 2018              12
## 4 Albany Albany County State Police 2018              12
## 5 Albany             Altamont Vg PD 2018              12
## 6 Albany          Bethlehem Town PD 2018              12
##   total_index_crimes violent murder forcible_rape robbery
## 1               3968     822     12            51     209
## 2                  1       0      0             0       0
## 3                130      16      0             6       0
## 4                158      30      0            25       0
## 5                  7       0      0             0       0
## 6                526      27      0             8       3
##   aggravated_assault property burglary larceny motor_vehicle_theft
## 1                550     3146      432    2584                 130
## 2                  0        1        0       1                   0
## 3                 10      114       19      82                  13
## 4                  5      128        5     121                   2
## 5                  0        7        1       6                   0
## 6                 16      499       28     466                   5
##              region
## 1 Non-New York City
## 2 Non-New York City
## 3 Non-New York City
## 4 Non-New York City
## 5 Non-New York City
## 6 Non-New York City
```

Data formatting and contingency tables - look by county, and then by agency in the worst and best counties.

```
totaled <- data %>% filter(agency != "County Total") %>%
  group_by(county) %>%
  summarize(vTotal = sum(as.numeric(violent)), pTotal = sum(as.numeric(property)), countyTotal = sum(as
  mutate(vProportion = vTotal/countyTotal, pProportion = pTotal/countyTotal)

orderedViolent <- totaled %>% arrange(desc(vProportion))
violentCounty <- data %>% filter(agency != "County Total") %>%
  filter(county == orderedViolent[1:5,]$county)
```

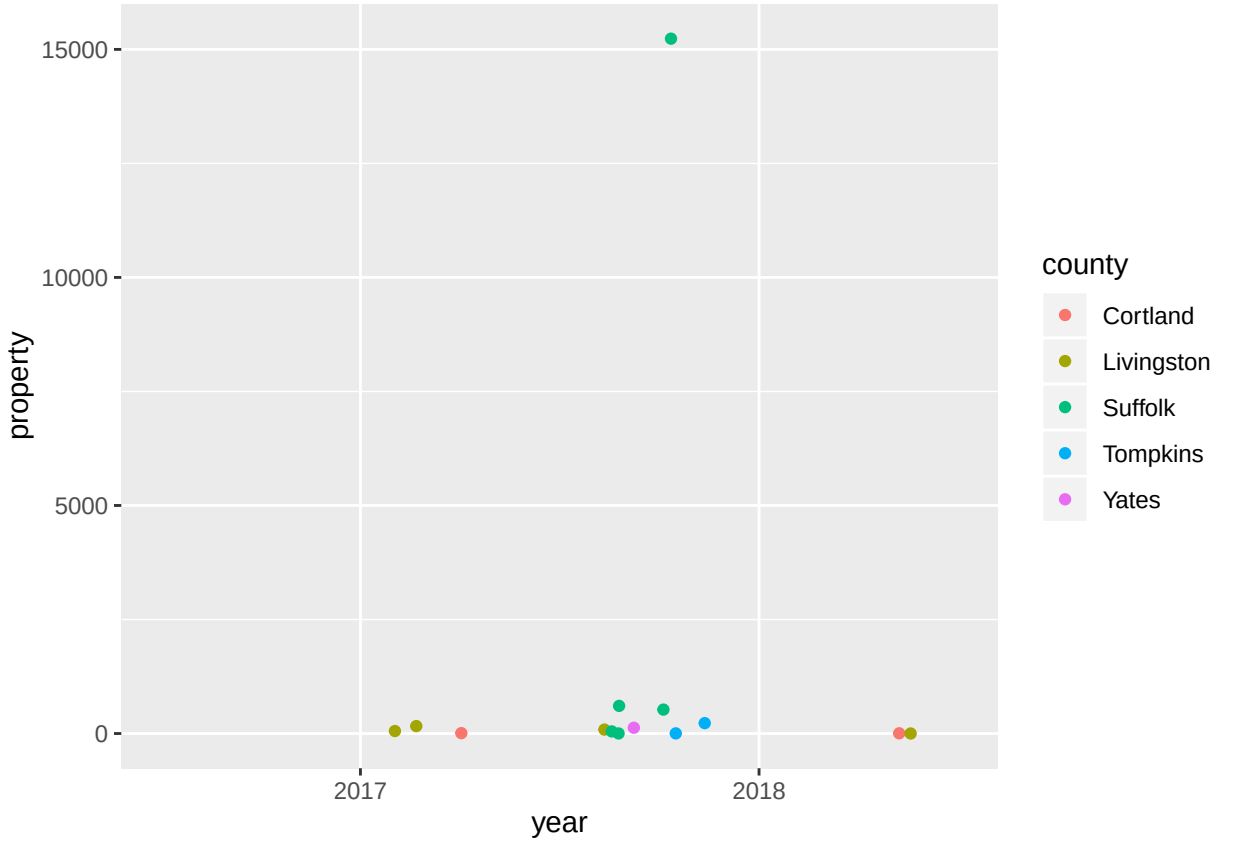
```
violentCounty$violent <- as.numeric(violentCounty$violent)
violentCounty$property <- as.numeric(violentCounty$property)

orderedProperty <- totaled %>% arrange(desc(pProportion))
propertyCounty <- data %>% filter(agency != "County Total") %>%
  filter(county == orderedProperty[1:5,]$county)
propertyCounty$property <- as.numeric(propertyCounty$property)
propertyCounty$violent <- as.numeric(propertyCounty$violent)
```

## Property Crime

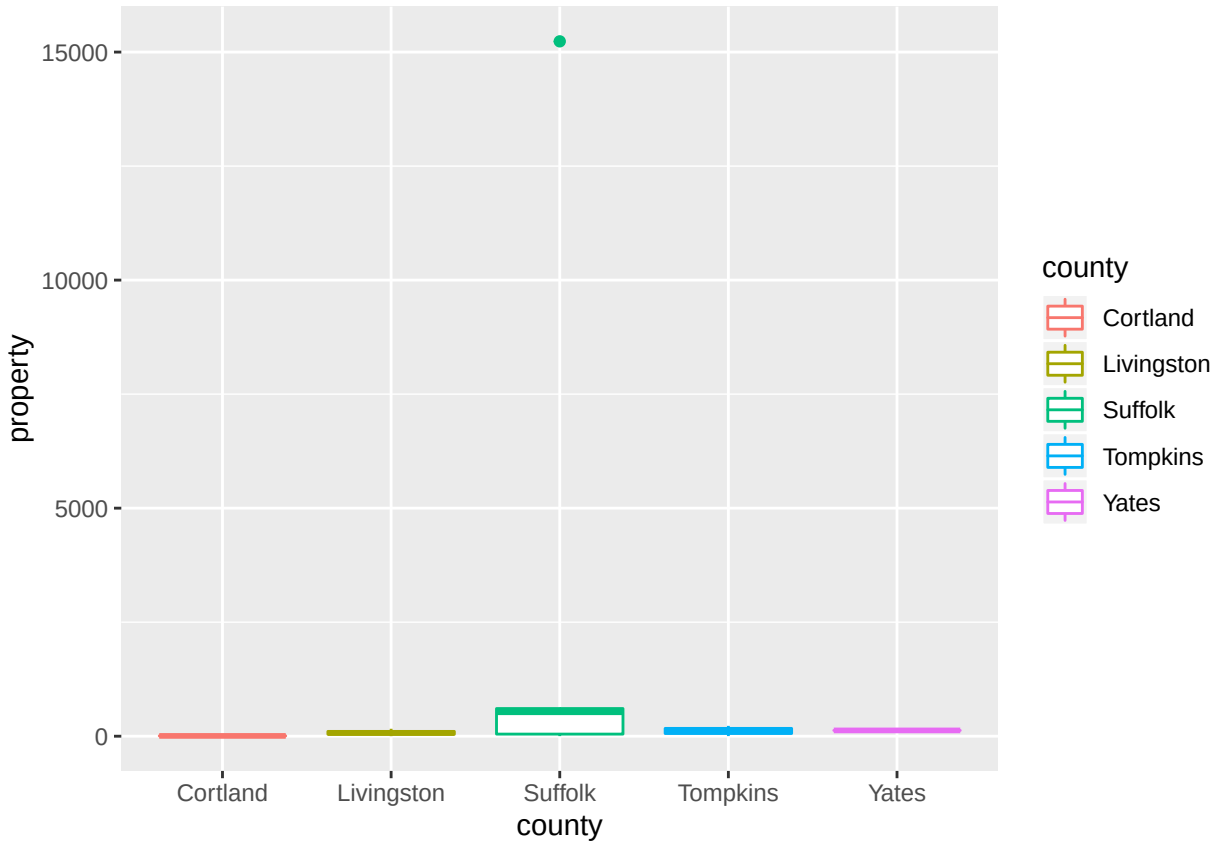
Let's look at plots for the top 5 property crime counties.

```
ggplot(propertyCounty, aes(year, property, color=county)) + geom_jitter()
```



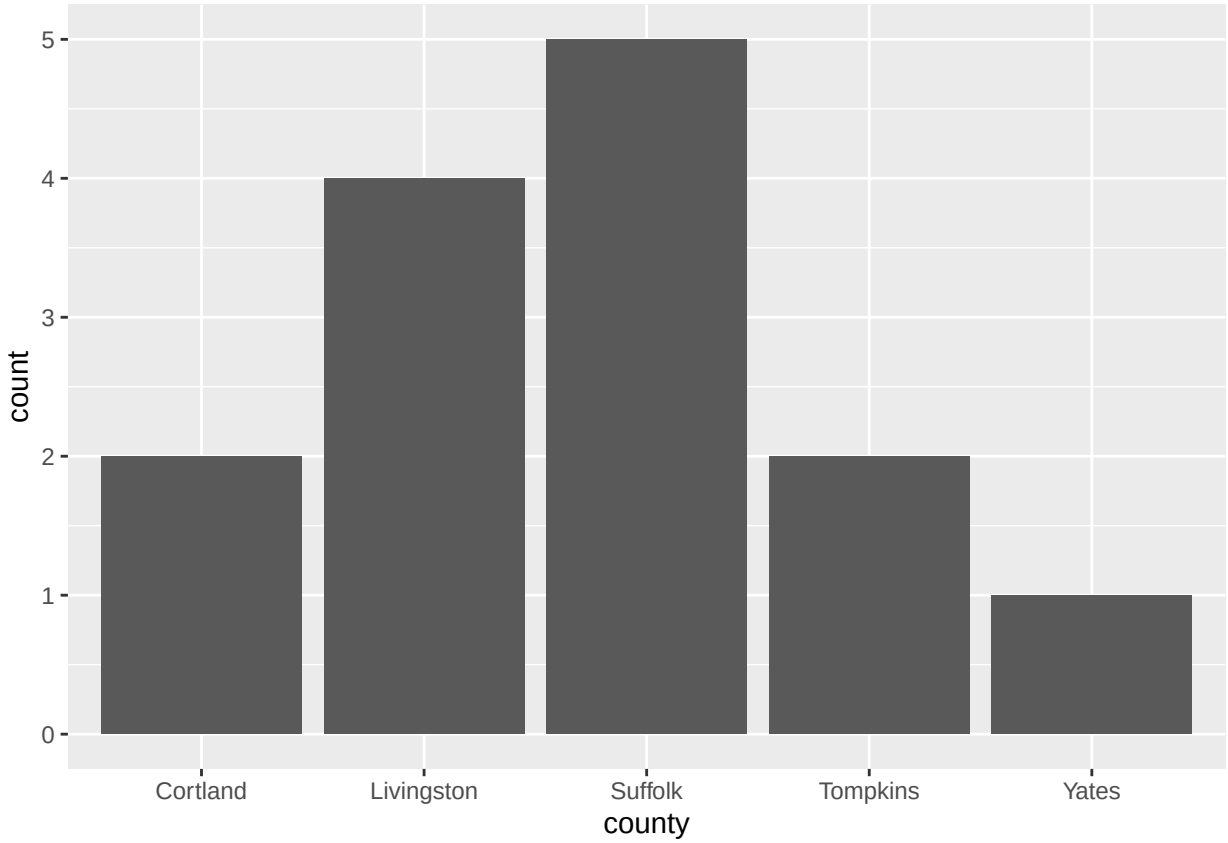

Suffolk looks like it has far more property crime than the other counties - let's look at a boxplot counties - essentially, looking at how the different agencies in these counties compare.

```
ggplot(propertyCounty, aes(county, property,color=county)) + geom_boxplot()
```



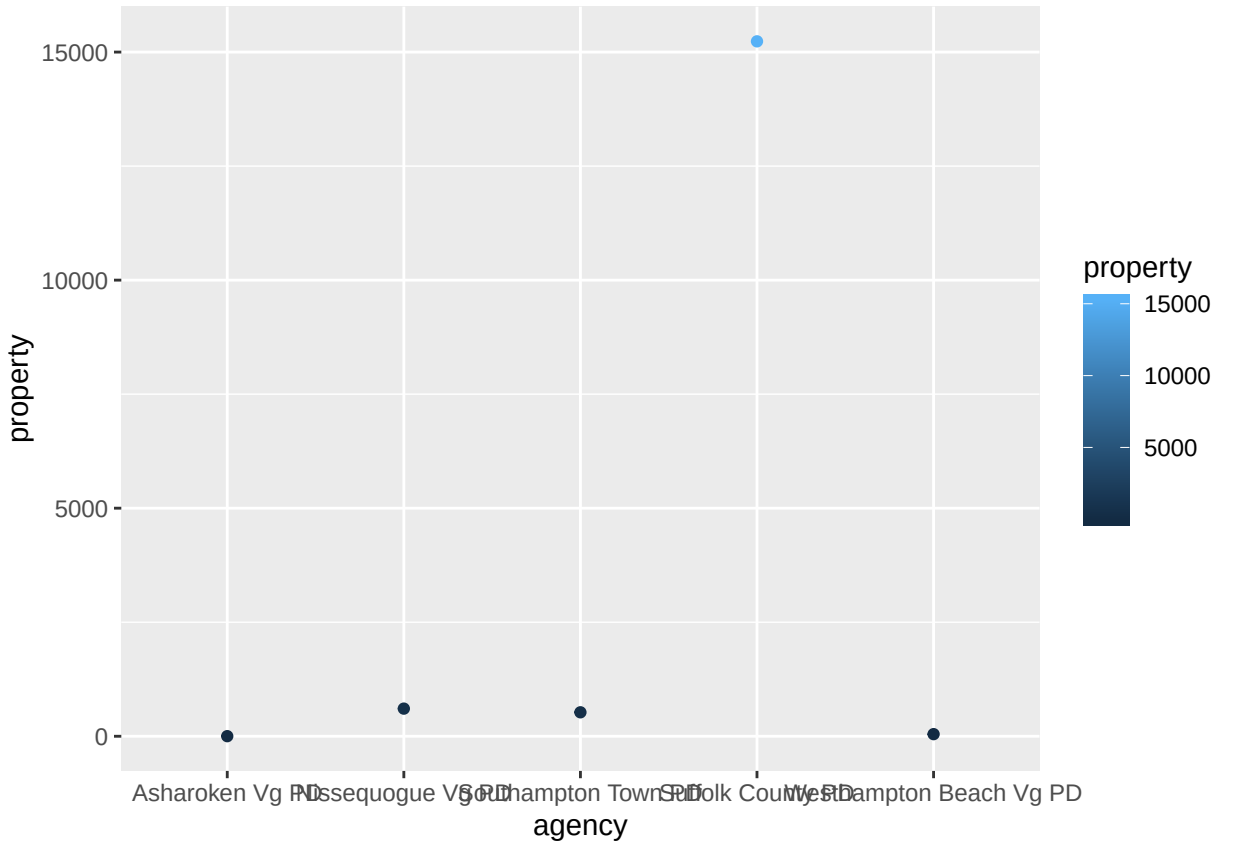

One nice thing to check here is also how many agencies there are in each of these counties.

```
ggplot(propertyCounty, aes(x=county, count=property)) + geom_bar()
```



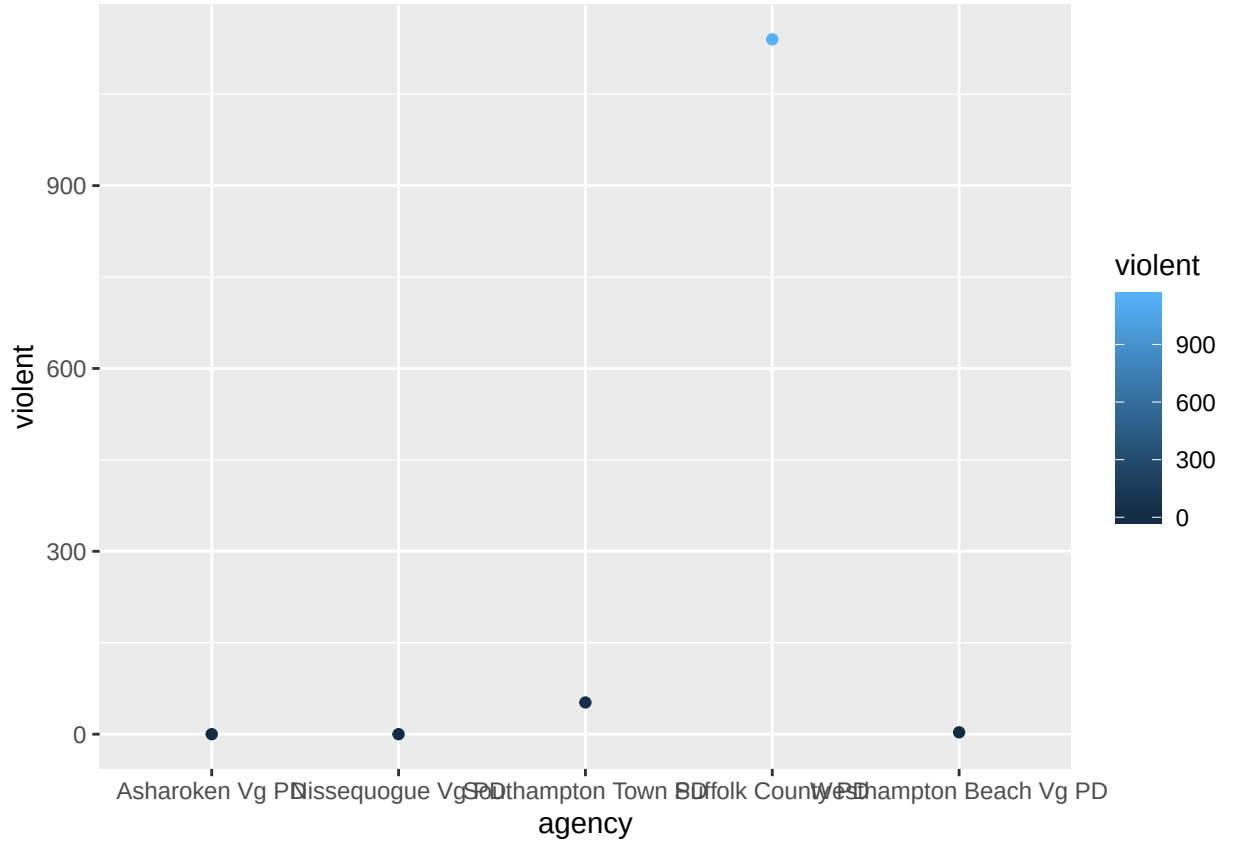

Suffolk does have the most agencies, but not by an absurd amount - there are other counties within 1 agency of it. Let's look at the agencies in Suffolk in more detail.

```
ggplot(propertyCounty %>% filter(county=="Suffolk"), aes(x=agency, y=property, color=property)) + geom_
```




Suffolk County PD has a LOT more property crime than the other agencies in Suffolk - how does violent crime compare?

```
ggplot(propertyCounty %>% filter(county=="Suffolk"), aes(x=agency, y=violent, color=violent)) + geom_po
```




Crime overall is much higher in Suffolk - it would be nice to see how that compares to the population in that area compared to the others.

Going back to an overall comparison of the counties, we can see that in the property category (where Sufolk was among the top 5), Sufolk has the most agencies. It would be nice to see proportionally how it stacks up to the other counties in terms of demographics/population as well.

Looking at a contingency table of these counties, we can also see that Sufolk only has data from 2018 - it would be interesting to track it and see how things change over time.

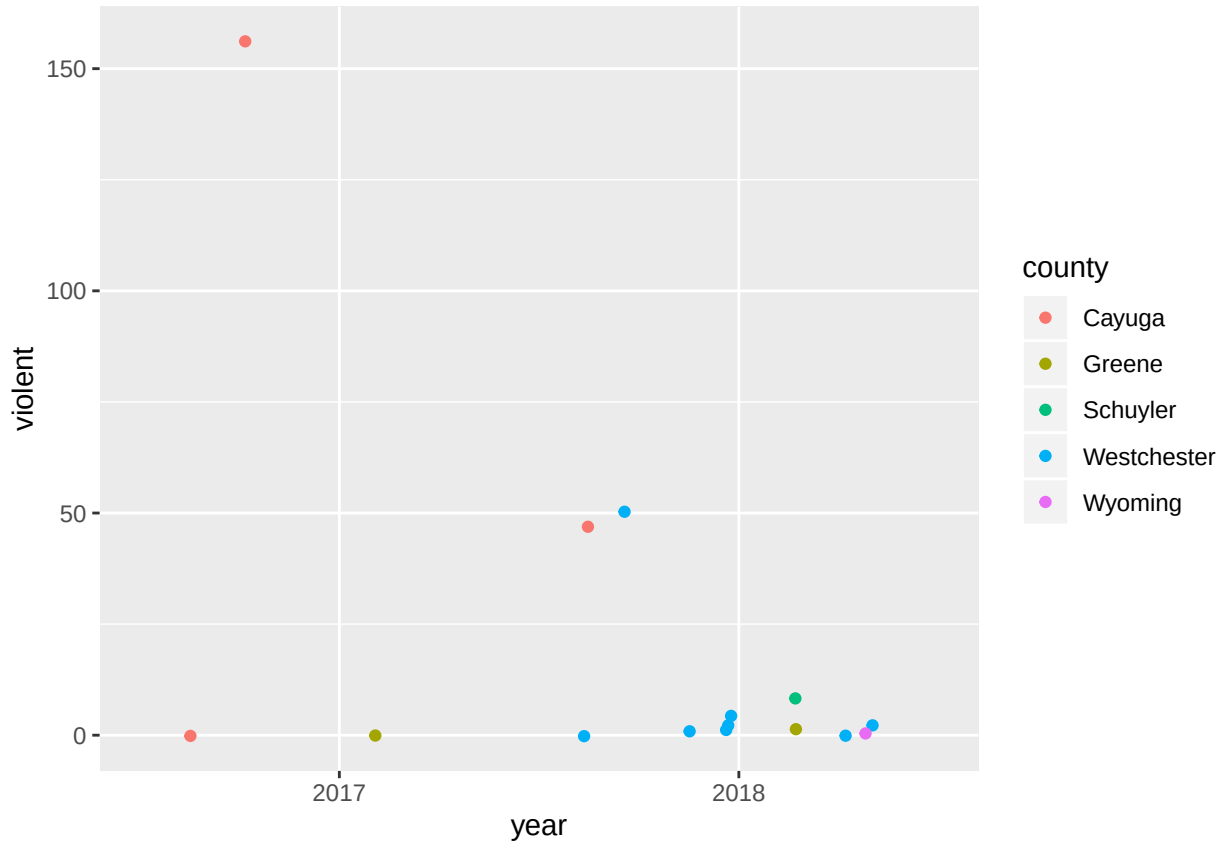
```
table(droplevels(propertyCounty$county), propertyCounty$year)
```

```
##
##              2017 2018
##   Cortland      1    1
##   Livingston    2    2
##   Suffolk       0    5
##   Tompkins      0    2
##   Yates         0    1
```

## Violent Crime

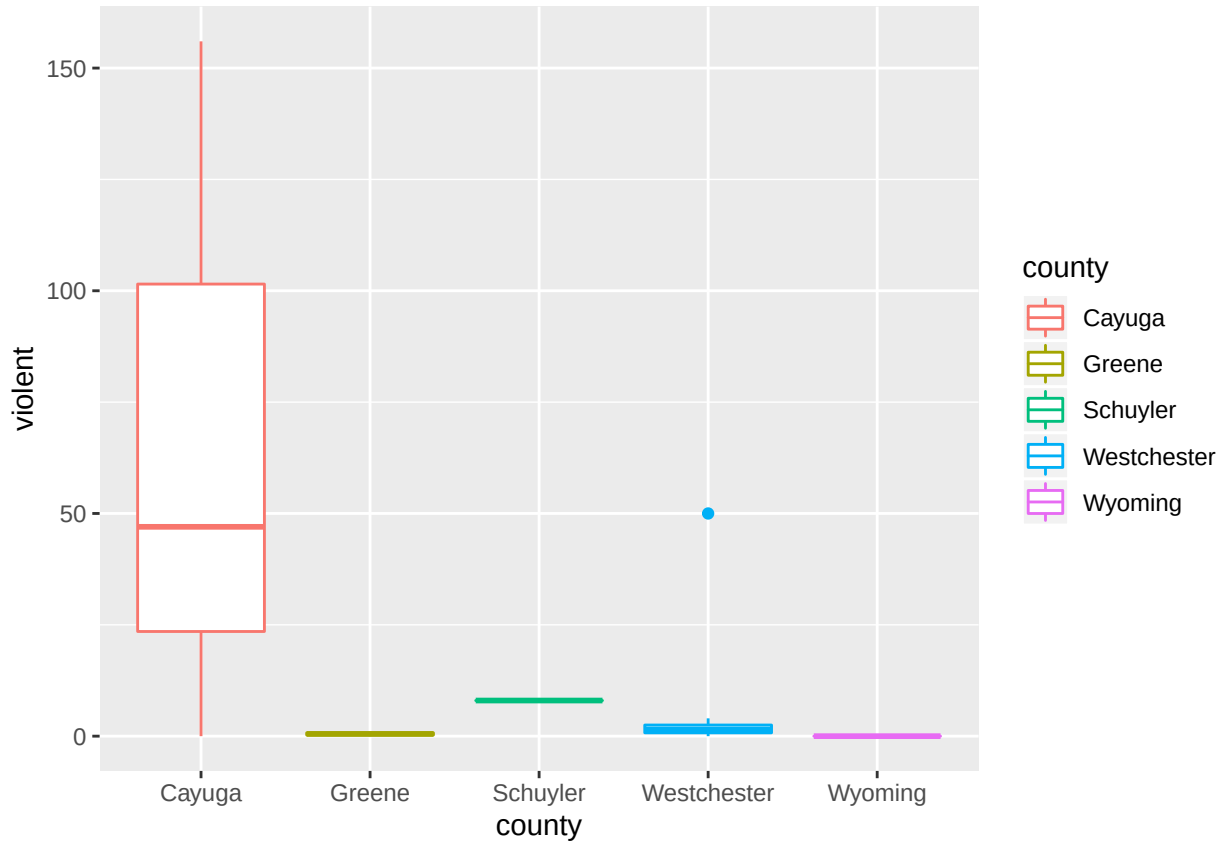
Now let's look at plots for the top 5 property crime counties.

```
ggplot(violentCounty, aes(year, violent, color=county)) + geom_jitter()
```



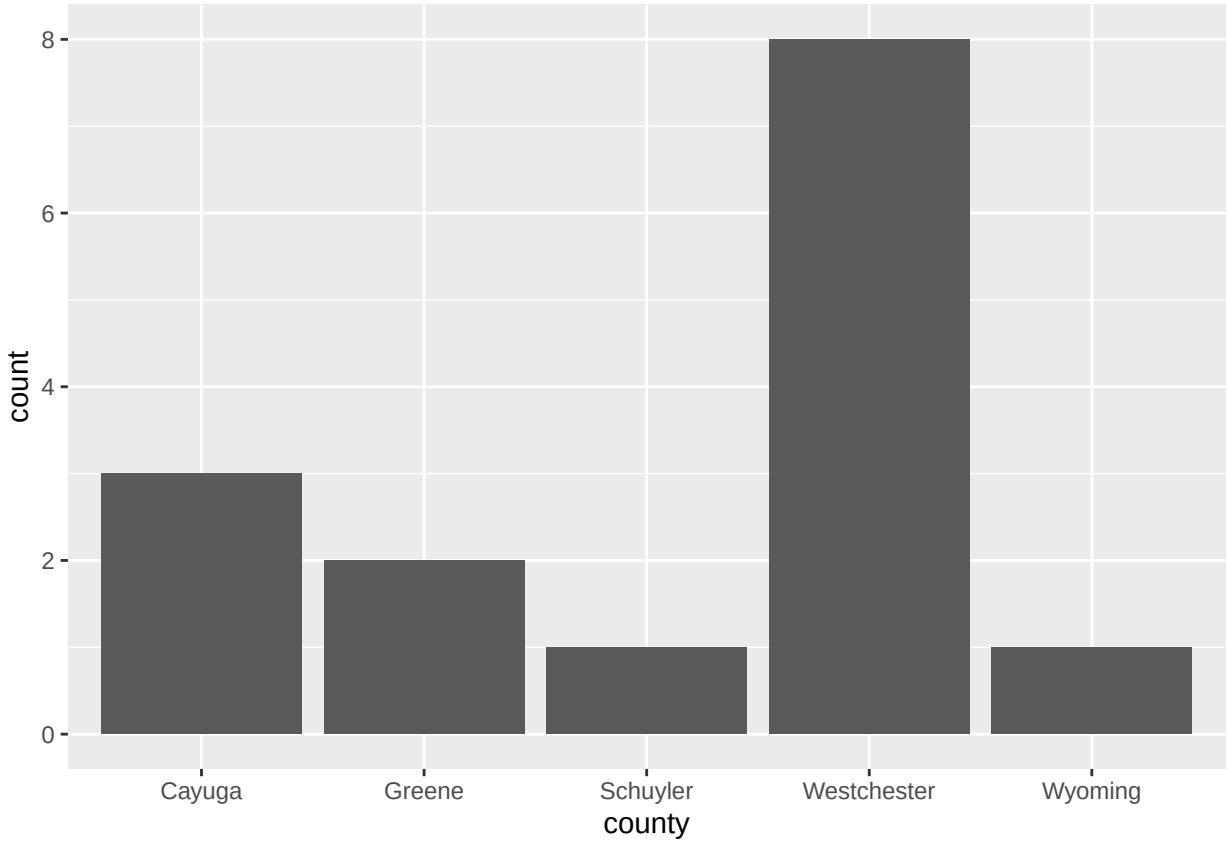

Cayuga had a lot of violent crime than the other counties, especially in 2017 - let's look at a boxplot of these counties, without taking year into account.

```
ggplot(violentCounty, aes(county, violent,color=county)) + geom_boxplot()
```



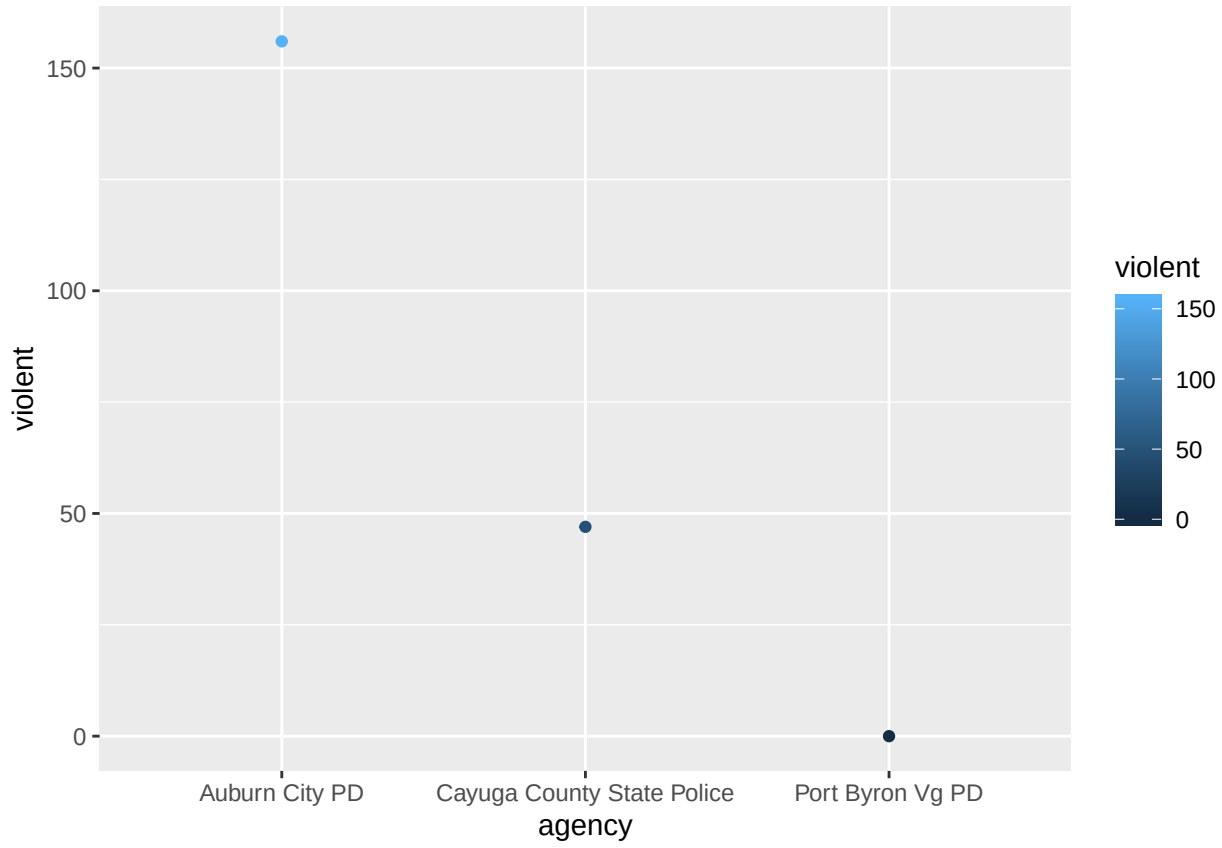

Cayuga has a huge spread here. Like with the property counties, let's look at the number of agencies in each county.

```
ggplot(violentCounty, aes(x=county, count=violent)) + geom_bar()
```




Cayuga doesn't have the most agencies, unlike Suffolk above. We definately want to look more into the agencies that may be seeing so much violence there.

```
ggplot(violentCounty %>% filter(county=="Cayuga"), aes(x=agency, y=violent, color=violent)) + geom_poin
```




Auburn City PD has a good amount of violence. Let's look at property crimes.

```
ggplot(violentCounty %>% filter(county=="Cayuga"), aes(x=agency, y=property, color=property)) + geom_po
```


Crime overall is much higher in the Auburn City PD agency than in the other agencies in Cayuga. As before, it would be nice to see how that compares to the population in that area compared to the others.

Looking at a contingency table of these counties, we can also see that Cayuga has less data in 2018, and Auburn City PD agency was not reported in 2018 - it would have been nice to see how crime changed from year to year there.

```
table(droplevels(violentCounty$county), violentCounty$year)
```

```
##
##                2017 2018
##   Cayuga          2    1
##   Greene          1    1
##   Schuyler        0    1
##   Westchester     0    8
##   Wyoming         0    1
```

```
cayuga <- violentCounty %>% filter(county=="Cayuga")
table(droplevels(cayuga$agency),cayuga$year)
```

```
##
##                              2017 2018
##   Auburn City PD                1    0
##   Cayuga County State Police    0    1
##   Port Byron Vg PD              1    0
```